The Journal of Ethics (2020) 24:401–425
https://doi.org/10.1007/s10892-020-09328-0

Beyond the Moral Influence Theory? A Critical Examination of Vargas's Agency Cultivation Model of Responsibility

Harry Harland[1]

Received: 21 November 2019 / Accepted: 15 February 2020 / Published online: 9 March 2020

Abstract

This paper repudiates Manuel Vargas's attempt to supplant the traditional moral influence theory of responsibility (MIT) with his 'agency cultivation model' (ACM). By focusing on fostering responsiveness to moral considerations, ACM purports to avoid the chief pitfalls of MIT. However, I contend that ACM is far less distinctive than it initially appears and so possesses all of MIT's defects. I also assail Vargas's counterfactual test for assessing whether a wrongdoer can respond to moral considerations. It is argued that the counterfactual test is epistemically redundant because it can only be fleshed out once we have settled the issues it is supposed to resolve. Moreover, it tacitly inverts the relation between freedom and responsibility—we cannot attribute free will to wrongdoers unless we have already established their blameworthiness. The upshot of this is that enquiries into an agent's freedom are irrelevant to the ethics of holding her responsible.

Keywords Free will · Moral responsibility · Agency cultivation model · Moral influence theory · Consequentialism · Moral agency · Manuel Vargas · Building Better Beings

In an ambitious work of impressive scope, Manuel Vargas espouses the 'agency cultivation model' of moral responsibility, which is intended to surpass traditional versions of the moral influence theory (Vargas 2013). This paper contends that Vargas's model is in fact indistinguishable from the theories it is supposed to supersede, leaving it vulnerable to the objections he hopes to avoid. In addition, I dispute his proposed model of free will on grounds of epistemic redundancy and conceptual distortion. Section 1 details Vargas's attempt to transcend the moral influence theory by employing the notion of responsiveness to moral considerations. Section 2 unpacks

Communicated by Neal Tognazzini.

✉ Harry Harland
hh441@cam.ac.uk

[1] Gonville and Caius College, University of Cambridge, Cambridge CB2 1TA, England, UK

his account of moral agency to reveal the strong affinities between the agency cultivation model and orthodox influence theories. Finally, Sect. 3 scrutinises the role that the concept of free will plays in Vargas's account. While he ostensibly treats free will as a precondition for moral responsibility, I argue that his theory only permits us to ascribe free will to wrongdoers if we have already ascertained that they are blameworthy. But once we have ascertained this, contemplation of a wrongdoer's freedom becomes superfluous—the concept of free will is left with no work to do. Furthermore, I contend that Vargas's abstruse counterfactual test for freedom represents a much more significant departure from folk morality than he acknowledges. It is suggested that the shortcomings of his conception of freedom arise because his methodological commitments pull him in different directions. On the one hand, he wants to tailor his metaphysical requirements to his goal of promoting moral agency. On the other, he seeks to preserve the intuition that we must be able to do otherwise in order to be responsible.

1 Beyond the Moral Influence Theory: The Agency Cultivation Model

In *Building Better Beings*, Vargas defends our 'responsibility-characteristic practices', which are roughly characterised by our propensity to express reactive attitudes in response to good or bad conduct (Strawson 1993; Vargas 2013: 1). His defence centres on the 'agency cultivation model' (ACM), which aims to incorporate insights from the moral influence theory as traditionally conceived (MIT) and reasons-responsive accounts of responsibility. Crudely stated, proponents of MIT hold that our responsibility practices are vindicated by their tendency to strengthen our dispositions to behave in morally desirable ways (see Nowell-Smith 1948; Smart 1961; Schlick 1962: ch 7; Dennett 2003, 2015; Arneson 2003; McGeer 2015). In contrast, reasons-responsive theories emphasise a connection between an agent's responsibility and her ability to act in accordance with reason (see Wolf 1990; Wallace 1994; Fischer and Ravizza 1998; Nelkin 2011; McKenna 2012). Drawing on each of these approaches, Vargas maintains that our responsibility practices are justified by their conduciveness to moral agency:

> When we hold one another responsible, we participate in a system of practices, attitudes, and judgments that support a special kind of self-governance, one whereby we recogni[s]e and suitably respond to moral considerations. So, roughly, morali[s]ed praise and blame are justified by their effects, that is, how they develop and sustain a valuable form of agency, one we ordinarily have reason to care about. (Vargas 2013: 2)

This focus on the impact of praise and blame means that ACM can straightforwardly be classified as a sort of moral influence theory, and Vargas himself claims to offer a rehabilitated version of MIT (Vargas 2013: 166). However, for ease of discussion, I shall continue to use 'MIT' to designate traditional moral influence theories and 'ACM' to designate Vargas's model (although we shall see that the two are far less distinct than they seem).

Vargas is drawn to MIT because it offers an appealing explanation as to why our responsibility practices are valuable: they improve our behaviour (Vargas 2013: 165–166). However, he also thinks it vulnerable to a number of challenges that ACM can avoid. For instance, because of its crude focus on positively influencing behaviour, MIT cannot delimit the range of responsible agents in an intuitively satisfying way. Since praise and blame can have desirable effects on infants and some non-human animals, MIT must classify such beings as responsible agents (Vargas 2013: 168). Campbell, for example, notes that '[i]t is quite possible, by punishing the dog who absconds with the succulent chops designed for its master's luncheon, favourably to influence its motives in respect of its future behaviour in like circumstances' (Campbell 1951: 447). However, it seems implausible to ascribe responsible agency to a dog. ACM supposedly allows us to avoid such untenable conclusions. Since infants and most non-human animals lack the faculties required to grasp moral considerations, it is inappropriate to subject them to the reactive attitudes. ACM thus 'permits us to draw the line in exactly the right place' (Vargas 2013: 188).

Another drawback of MIT is that only committed consequentialists are likely to find it persuasive (Vargas 2013: 171). Indeed, a similar problem befalls most ethically-grounded theories of responsibility: 'when philosophers have worn their commitments to some or another normative ethical theory on their sleeve, there tends to be comparatively little uptake of that work internal to the subfield concerned with free will and moral responsibility' (Vargas 2013: 128, fn29). Thus, rather than adopting a 'partisan ethics' approach, Vargas claims to embrace a 'limited ethics' methodology (Vargas 2013: 126–130). He states that

> [G]iven the contentious nature of normative ethics, a theorist of responsibility should treat it as a desideratum that any proposed account of moral responsibility be somewhat insulated from commitments to a specific theory of normative ethics. If we accept this methodological constraint, moral influence theories are problematic not because of consequentialism as such, but because they imply a commitment to a specific theory of normative ethics. (Vargas 2013: 171)

Vargas is happy to concede that ACM is consequentialist with respect to its focus on outcomes. However, he thinks that one need not be a wholehearted consequentialist in order to embrace it. For instance, a deontologist might maintain that praise and blame are justified because they increase compliance with deontological norms. Unlike MIT, then, ACM is only 'modestly teleological' (Vargas 2013: 173). As Vargas puts it:

> The present account should be understood as *modular*. That is, it can be integrated with different ethical theories without affecting the basic justification for the distinctive norms of responsibility. However, how we understand various elements of the account, and what external constraints structure and limit the account will vary by pairing. So, for example, if paired with consequentialism, we should look to the consequentialist theory of the good to inform the account of moral considerations. If paired with a Kantian ethical theory, moral considerations will presumably be understood in a different way, connected to

> the Categorical Imperative. Since the moral influence theory is not intended to be an account of right action, but rather a broadly modular account of moral responsibility, details about how to understand the content of moral notions invoked by the account will be subject to variation. (Vargas 2013: 185–186, emphasis in original, footnote omitted)

Vargas rightly notes that the pursuit of ACM's goals can be constrained by nonconsequentialist precepts (Vargas 2013: 191–193). For instance, if we wanted to incline a wrongdoer away from misconduct, we would enjoy greater success by extending our resentment to his children and other relatives.[1] But this could still be deemed unjust if we embraced a principle of fairness that precluded blaming the innocent. However, we could just as easily append deontological qualifications to any version of MIT. Thus, these considerations give us little reason to prefer ACM's agential focus to the dispositional focus of MIT. ACM and MIT are equally compatible with deontological side constraints.

A further shortcoming of MIT is that it cannot distinguish between genuine blame and blame that is feigned for the purpose of improving its target (Vargas 2013: 168–169). Consider, for example, how our reactions would differ if our precious wallpaper was defaced by either a spiteful houseguest or a creative toddler. While full-blooded blame would befit the houseguest, our blame toward the toddler ought to be half-hearted at most. However, MIT disallows distinctions of this sort. If treating children severely produces desirable effects, then we should treat them severely. Vargas thinks ACM can escape this concern. He maintains that genuine (full-blooded) expressions of blame have an important cognitive element. They involve the judgment that their object satisfies the criteria for responsible agency. This observation cannot aid MIT because MIT will ascribe agency to many small children and non-human animals, as mentioned earlier. ACM, on the other hand, excludes such beings from the category of moral agents due to their moral and intellectual limitations (Vargas 2013: 188–190).

Another powerful objection to MIT (which Vargas does not make explicit) is that its rationale countenances odious forms of manipulation. If we view moral improvement as a desideratum, we must find some principled way to permit praise and blame while disallowing more sinister influences. For example, Herbert Morris famously recoiled at rehabilitative theories of punishment because of their consistency with 'the giving of an evil-tasting pill' to produce 'instantaneous truth or aversion to acting violently' (Morris 1981: 265–266). It is not immediately obvious that ACM can address such concerns. After all, it is conceivable that we may invent evil-tasting pills or neurosurgical techniques that generate moral epiphanies or give us the will power required to act on our moral values. Perhaps one day we could induce powerful hallucinations that provide a vivid awareness of what it feels like to be a victim of one's crimes (Dolinko 1999: 359). In response, Vargas would no doubt point out that deontological side constraints could prohibit such practices. However, this retort

[1] I have borrowed this example from Boonin (2008: ch 2).

could equally well be made by proponents of MIT. Moreover, the tenability of this response is highly questionable, as will become clear in Sect. 2.

Vargas considers three more objections to MIT that I will not dwell on at any length. This is because Vargas's rejoinders to them can just as easily be made by supporters of MIT. (Indeed, Vargas himself does not offer these as reasons to prefer ACM over MIT.) My chief purpose in the next section is to repudiate the claim that ACM represents a meaningful departure from MIT. Since the following rejoinders do not purport to establish ACM's superiority, they can be dispensed with briefly.

First, Vargas notes that MIT is often alleged to conflate 'being responsible with judgments about the appropriateness of holding responsible' (Vargas 2013: 168). According to MIT, an agent is responsible just in case praising or blaming her will produce certain desirable effects. But, the objection runs, an agent can be responsible even though it would be inappropriate to hold her responsible. For instance, it could be improper to hold somebody responsible for her first impolitic remark, even though she is responsible for that remark (Vargas 2013: 169). Like proponents of MIT, Vargas accepts that influenceability is the mark of responsibility, but notes that factors external to ACM, such as 'considerations of justice, benevolence, [and] prudence' could render it inappropriate to hold someone responsible (Vargas 2013: 190). Given the discussions in the last few paragraphs, it should be immediately obvious that this line of reasoning could also be adopted by MIT's defenders.

The second objection is that MIT does not account for the phenomenology of holding others responsible (Vargas 2013: 169). As PF Strawson notes, the reactive attitudes express 'how much we actually mind, how much it matters to us, whether the actions of other people—and particularly of some other people—reflect attitudes towards us of goodwill, affection, or esteem on the one hand or contempt, indifference, or malevolence on the other' (Strawson 1993: 49). We thump the table with anger when a colleague will not repay a debt; we express teary-eyed gratitude to friends who console us in times of need. But MIT merely recommends that we behave *as though* we feel such emotions—it is indifferent about whether we actually experience them. It may, for instance, commend the outward appearance of outrage when one is really in a state of perfect equanimity. In other words, MIT cannot explain why genuine reactive attitudes are preferable to hollow imitations. In response to this concern, Vargas observes that most instances of praise and blame reflect genuine emotions and he is thankful this is so. After all, if we were not 'interpersonally engaged', our inclinations to hold others accountable would decrease markedly (Vargas 2013: 193). Moreover, ACM does not prescribe that each instance of praise and blame must be informed by cold, forward-looking calculations. If the interpersonal engagements that drive our responsibility practices are generally efficacious (that is, if they are sufficiently conducive to moral considerations responsiveness), then we are permitted to preserve them without having to reflect much about their utility in particular instances (Vargas 2013: 193–194). This line of reasoning, whatever its merits, is clearly compatible with MIT and so I will say nothing more of it here.

Finally, the third objection is that MIT's forward-looking focus mischaracterises our responsibility practices (Vargas 2013: 169–170). Consider, for example, the 'Butcher of Lyon' Klaus Barbie, a Nazi war criminal who was not convicted until 1987. Upon

learning of how Barbie tortured children and flayed people alive, nearly all of us will feel indignation toward him. But our indignation is for what *he has done* rather than what *he will do*. We might hope that our opprobrium will deter would-be war criminals, but this is not our reason for expressing it. Some ascriptions of responsibility are, as Gerald Dworkin remarks, 'not oriented toward the future but are, so to speak, for the record' (Dworkin 1986: 424). Vargas's response to this issue overlaps significantly with his response to the previous one. He begins by endorsing Strawson's observation that the reactive attitudes reflect judgments about the personal qualities others have exhibited in their actions. With this in mind, Vargas claims:

> Assuming a good will is at least sometimes reflective of moral considerations, it is reasonable to think that learning to track a good will can play a role in learning to track moral considerations. Perhaps more importantly, our reactions of gratitude can signal that we recogni[s]e that other agents are responding to what we regard as appropriately agency-guiding considerations. Of course, sometimes these considerations are extra- or non-moral, but inasmuch as gratitude reliably reflects appreciation of moral considerations-governed agency too, gratitude has all the license we can hope for. Similar remarks hold for other backward-looking attitudes: as long as they plausibly play a role in the social and intrapersonal economy of governance by moral considerations, there is no objection here. (Vargas 2013: 194)

In other words, backward-looking ascriptions of responsibility are justifiable because they can further the moral development of both oneself and others. Like Vargas's previous riposte, it is obvious that this response can be made in defence of MIT. However, a couple of things are worth noting. First, for someone trying to make friends in diverse quarters, Vargas's claim that a future-oriented justification of our practices is all 'we can hope for' is sweepingly dismissive. Second, this claim is worlds apart from one of his opening assertions that 'we are beings for whom morali[s]ed praise and blame make sense, *partly* because of the effects of praise and blame' (Vargas 2013: 3–4, emphasis added). This initial suggestion of a rich justification that draws only in part on consequentialist considerations belies the subsequent claim that a future-oriented rationale is the only game in town.

Although some will be repelled by Vargas's heavy reliance on forward-looking concerns, ACM may appear to constitute a significant advancement beyond MIT. Indeed, one reviewer commends his 'admirable job of showing how [ACM] is largely immune to the sorts of worries thought to plague [MIT]' (Capes 2016: 248). However, such praise is misplaced. Once we have unpacked Vargas's conception of moral agency in the next section, we shall see that it is no different to MIT's notion of being disposed to behave morally.

2 Responsiveness to Moral Considerations: A Distinction Without a Difference

In order to fully appreciate ACM's failure to move beyond MIT, we should first familiarise ourselves with Vargas's account of how praise and blame can promote moral agency. He begins by noting that norms are often internalised as a result of sustained exposure to external reasons for compliance (Vargas 2013: 175–177). The reactive attitudes and their associated behaviours 'initially work by providing external motivation for agents to track moral considerations and regulate their behavio[u]r in light of them' (Vargas 2013: 175).[2] However, 'norms for which we start off having only external motivations to obey will, under many conditions, go on to become internali[s]ed. When that happens, the norms are experienced as intrinsically motivating' (Vargas 2013: 175). External motivation is left undefined but we might stipulate the following:

> **External Motivation** A person is externally motivated to ϕ (where 'ϕ' represents adherence to some norm) to the extent that: (1) he is in fact motivated to ϕ; and (2) his motivation to ϕ is dependent upon his expectation that certain beneficial or adverse personal consequences will respectively follow from his either ϕing or not ϕing.

For instance, an atheist politician may be motivated to attend church in order to win religious votes. Similarly, a misanthropic dog walker may be motivated to pick up her pet's droppings in order to avoid a fine. The politician expects beneficial consequences from compliance; the dog walker expects adverse consequences from non-compliance. *Intrinsic* motivation seems to be the inverse of external motivation: non-prudential motivation to comply with some norm. Vargas emphasises that 'norms of praise and blame can come to structure the deliberations of agents even when actual expressions of praise and blame are unlikely or absent' (Vargas 2013: 175). He also cites a well-known piece about internalisation by Sripada and Stich (2006). These authors state that 'according to the internali[s]ation hypothesis, individuals exhibit a characteristic style of motivation in which the individual intrinsically values compliance with moral rules even when there is no possibility of sanction from an external source' (Sripada and Stich 2006: 285–286). They also note that intrinsically motivated people are 'disposed to comply with norms even when there is little prospect for instrumental gain, future reciprocation, or enhanced reputation, and when the chance of being detected for failing to comply with the norm is very small' (Sripada and Stich 2006: 285). Hence, it seems that an individual will have internalised some norm to the extent that he is intrinsically motivated to comply with that norm. And with respect to intrinsic motivation, we might stipulate the following:

[2] What Vargas should be saying here is that praise and blame provide external motivation for tracking *other people's convictions* about moral considerations. This slip is corrected later, where he states that 'morali[s]ed blame tends to push the target's attention to considerations that *others perceive as morally salient*' (Vargas 2013: 198, emphasis added).

> **Intrinsic Motivation** A person is intrinsically motivated to ϕ (where 'ϕ' represents adherence to some norm) to the extent that: (1) he is in fact motivated to ϕ; and (2) his motivation to ϕ is *not* dependent upon his expectation that certain beneficial or adverse personal consequences will respectively follow from his either ϕing or not ϕing.

There are obviously multiple routes to intrinsic motivation. One person may be intrinsically motivated to ϕ as a result of sustained moral reflection, another as a result of unthinking conventionalism. The important point is that the motivation is not presently tied to external incentives (though it may have been at one point). None of this entails that an agent cannot be both intrinsically motivated to ϕ and externally motivated to ϕ (Sripada and Stich 2006: 285). For example, our dog walker could be motivated by both other-regarding and prudential concerns, and each motivation may be individually sufficient to lead to action. But how does Vargas connect intrinsic motivation and responsiveness to moral considerations? I shall encapsulate that connection in one final stipulation, in support of which I will adduce several passages from Vargas.

> **Responsiveness to Moral Considerations** A person is responsive to some moral consideration *M* to the extent that he is intrinsically motivated to act in accordance with *M*.

Curiously, Vargas immediately abandons the language of intrinsic motivation after his brief discussion of norm internalisation, and he mentions internalisation only once more—in a subsequent chapter where he recommends adjusting our responsibility practices in light of 'the need of agents to internali[s]e norms of action for moral considerations' (Vargas 2013: 214). He states that a moral consideration is 'a consideration with moral significance such that, were one to deliberate about what to do, it ought to play a role in those deliberations' (Vargas 2013: 203). Talk of *considerations* is preferred over that of *reasons* because it avoids 'the suggestion of an account of responsible agency that is particularly bound up in a rationalistic conception of agency' (Vargas 2013: 203). This is not to deny the moral agency of Spock-like figures who act on the basis of a cool and rational perception of moral reasons. Rather, it is to affirm the moral agency of those who act in a far less deliberative manner. As Vargas states: 'What I have in mind by moral considerations includes things that we recogni[s]e as reason[s] in as fully rationalistic a sense as you like, but also things that are largely or perhaps exclusively affective' (Vargas 2013: 203, fn6). At later junctures, we find two more important passages:

> [S]ensitivity to moral considerations may involve susceptibility to a nagging feeling, reacting to a dim hope, being able to imagine the situation of another, or attending to an inarticulate, largely inchoate suspicion about things. Just how any of this happens may vary from person to person and from situation to situation. (Vargas 2013: 208–209)

And:

> [D]etection of moral considerations need not be conscious, and the agent need not recogni[s]e that it is a moral consideration *qua* moral consideration that is moving him or her to act. Detection of moral considerations amounts to awareness of moral considerations, and that awareness need not be conscious or even necessarily explicable by the agent. (Vargas 2013: 217)

These remarks make it clear that Vargas embraces a highly capacious understanding of responsiveness to moral considerations. Indeed, it even allows us to describe some moral sceptics as morally responsive. Consider, for example, a philanthropic nihilist whose charitable endeavours result from an inner drive to help the unfortunate. Clearly, she would reject any claim that she is motivated by moral concern. However, for Vargas, there is no cognitive requirement that she construe her behaviour as moral in nature. It is enough that she is (unwittingly) acting in accordance with a moral standard on the basis of a 'nagging feeling'. After all, moral considerations-responsiveness can be 'exclusively affective'. These reflections make it difficult to see any difference between being *intrinsically motivated* to act in accordance with some moral consideration and being *responsive* to that moral consideration. Moreover, Vargas's lengthy discussion of how praise and blame promote norm internalisation would be superfluous if he did not see an important connection between intrinsic motivation and moral considerations-responsiveness. Section 2.2 will argue that Vargas's conception of moral agency is so commodious that his agency cultivation model is indistinguishable from the moral influence theory. But first, let us query whether paradigmatic examples of ACM in action can really be said to foster moral agency.

2.1 Taking the 'Agency' out of Moral Agency

Contemplation of two hypotheticals allows us to see that ACM is not really about the promotion of moral agency. First, imagine the following scenario:

> John is born into a religiously conservative society that strongly denounces homosexuality. During his adolescence, he begins to experience homosexual urges. However, whenever such impulses arise, he is struck by fears of hellfire and ostracism. As a result, John develops an aversion so visceral that he never acts on his homosexual desires. With time, his urges no longer evoke terror, but the aversion remains powerful. Indeed, John comes to regard homosexuality as he does incest—as something he *just knows* to be deeply immoral. Although John is strongly committed to this view, he lacks the reflective faculties required to provide a compelling rationale for it. When asked to why homosexuality is wrong, he merely falls into the tautology of replying: 'It just is!'

What, on Vargas's account, are we to make of John's development? Has his society cultivated within him a 'valuable form of agency'? (Vargas 2013: 2) Initially, his motivation to comply with the norm precluding homosexuality is entirely external. He is driven only by a fear of suffering unpalatable personal consequences (hellfire and ostracism). But his motivation eventually becomes intrinsic: even when external

reasons for compliance do not present themselves, John remains robustly indisposed to perform homosexual acts. Further, his aversion translates into a moral belief that homosexual acts are wicked. He may find this belief scarcely explicable, but, as we have seen, Vargas does not place a strong cognitive requirement on moral agency. Must Vargas conclude that John's responsiveness to moral considerations has been enhanced? Of course not—the norm John has internalised carries no genuine moral force. Consensual intercourse between same-sex adults is morally permissible. But now consider a parallel scenario:

> Fred is born into a liberal society that detests racism. At an early age, he experiences racist impulses which manifest themselves through cruel and derogatory behaviour. Such behaviour brings about various personal costs for Fred. His parents confiscate his tablet computer; his teacher secludes him in the naughty corner; his friends are not allowed to play with him. Initially, his prejudice remains undiminished but Fred certainly learns to think twice before giving expression to it. After several years of self-policing, he develops a deep antipathy toward his intolerance. When prejudicial sentiments enter his mind, he no longer imagines the cost of acting on them but he still feels a strong urge to quash them. This aversion convinces Fred of the iniquity of racism. But when asked why members of other races should be treated equally, he can say little more than 'They just should!'

In Fred's case, a genuine moral norm has been internalised, and so Vargas may be willing to concede that his moral agency has been enhanced. However, we should be reluctant to take this view. Since Fred's process of internalisation was analogous to John's, Fred remains unworthy of the label *moral agent*. The views of both men are not the products of careful reflection; instead, they result from unthinking capitulation to social pressures. Fred, of course, at least had the good fortune to swallow a true belief and we may very well be glad of this. However, under different circumstances, he could come to endorse far more odious norms. After a few years with the Westboro Baptist Church, we might find him chanting vulgar mantras while picketing the funerals of fallen soldiers. Let us suppose that almost all of the moral precepts Fred endorses are correct and have been internalised through similar social processes. Because he is unwaveringly disposed to act morally, Vargas must now view Fred as a paradigm of moral agency. However, the opposite conclusion is more apposite. There is an inverse relationship between moral agency and the susceptibility of one's moral convictions to external forces. Fred is not a moral agent; he is a moral drifter—a mere receptacle of social mores. The fact his moral convictions happen to be true does not make him a moral agent.

To avoid this line of criticism, Vargas might append stronger cognitive requirements to his conception of agency. For example, he could take Ronald Dworkin's view that moral agency involves subjecting one's moral convictions to careful scrutiny in light of one's other moral convictions, empirical beliefs, the requirements of logic, and so on (Dworkin 2011: ch 6). However, this variety of agency is highly demanding. It requires mental fortitude, intelligence, and creativity. And we cannot take it for granted that these characteristics could be enhanced meaningfully by exposure to praise and blame. Hence, if Vargas adopted a Dworkinian conception

of agency, his claim that our responsibility practices are effective at fostering moral agency would lose much of its plausibility. At the very least, it would require vastly greater empirical substantiation than he provides. He therefore appears to face the following problem: the richer the conception of agency employed, the less tenable ACM becomes. The upshot of this is not that we should withhold praise and blame from people like John and Fred—there are obviously good reasons for holding them accountable. However, we should not conclude that doing so promotes their moral agency. In other words, *agency cultivation model* is a misnomer. But it is no innocuous misnomer. This is not an issue of mere semantics. After all, it is Vargas's talk of moral agency which gives plausibility to his claim to have transcended traditional versions of the moral influence theory. Now that we have unpacked his account, we are well poised to expose the falsity of this claim.

2.2 The Objections to MIT Revisited

Recall the first objection discussed in Sect. 1: namely, that MIT cannot delimit the range of responsible agents in an intuitively satisfying manner. Vargas uses the language of moral agency to deflect this objection. He tells us that his focus on agency allows him to 'draw the line [between responsible and non-responsible agents] in exactly the right place' (Vargas 2013: 188). However, his conception of moral agency is so diluted that ACM would draw the line exactly where MIT does. For Vargas, responding to moral considerations amounts to little more than acting morally without regard to prudential concerns. Indeed, as an example of responding to moral requirements, Vargas cites '[b]eing able to perceive when a quiet friend is in need of consolation' (Vargas 2013: 208). However, this ability is possessed by many small children and non-human animals, as all parents and dog-owners know. Indeed, the renowned primatologist Jane Goodall has observed chimpanzees and bonobos soothing defeated friends, punishing delinquents, and adopting orphaned infants (Goodall 2010). Even a rat will liberate and share food with a caged companion (Bartal et al. 2011).[3] Infants aged between 18 and 30 months can respond to distress in others by sharing toys, providing hugs, and seeking the help of adults (Zahn-Waxler et al. 1979). Thus, contrary to what Vargas claims, ACM will in fact ascribe responsible agency to beings that he considers intuitively unresponsible. Far from establishing that ACM delimits the range of responsible agents in a more satisfactory manner than does MIT, Vargas fails to establish that ACM delimits that range differently at all.

Because the respective classes of responsible agents under MIT and ACM are more or less coextensive, Vargas's attempt to distinguish moral and non-moral forms of influence is also unpersuasive. Vargas, it should be remembered, claims that wholehearted ascriptions of responsibility are reserved only for those we judge to fall within the class of responsible agents. However, by inadvertently locating many children and non-human animals within that class, he has committed himself

[3] I owe this reference and the previous one to Churchland (2019: 166).

to the view that such beings are in fact apt candidates of wholehearted resentment and indignation. A toddler might not be a Dworkinian agent, but we can surely disincline her from drawing on the walls through stern disapprobation. Vargas has not explained why ACM provides grounds for mollifying our responses to beings who appear to lack full responsibility.

Now let us turn to MIT's 'partisan ethics' problem. While only consequentialists are drawn to MIT, Vargas thinks theorists of sundry ethical commitments can readily embrace ACM. One reason for this is that ACM can be combined with deontological precepts that constrain its pursuit of 'moral agency'. In Sect. 1, it was quickly pointed out that this is also true of MIT. I also promised to show that even a side-constrained version of ACM would not be agreeable to most nonconsequentialists. Before supporting this claim, I wish to defend Vargas against the objection that a society which adhered to ACM would violate the second formulation of Kant's Categorical Imperative: 'So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means' (Kant 1998: 38, emphasis removed). Nadine Elzein makes this complaint, stating that 'it is not easy to justify Vargas's [account] without abandoning the Kantian principle that prohibits treating any person merely as a means' (Elzein 2013: 212). This charge is often made against the deterrence-oriented rationale for punishment. Despite being a staunch deontologist, Matthew Kramer has defended the deterrence-oriented rationale for capital punishment against such allegations (Kramer 2011: 29–30).[4] His reasoning is echoed in my rebuttal of Elzein.

There are obvious reasons to think that ACM contravenes Kant's second formulation. As Elzein points out, ACM recommends that we impose unpleasant experiences on people for the benefit of others. It tells us to blame a wrongdoer so that society profits from his moral improvement. But in doing so, we would be using him as a mere instrument in the pursuit of the public good (Elzein 2013: 221–223). It is indisputable that ACM offers instrumental reasons to hold others responsible. However, it would be a mistake to think that any responsibility system informed by ACM would use people *solely* as a means.

ACM is immune to the *mere means* objection because it offers a system-level justification of our responsibility practices. It seeks to propagate responsibility norms that strengthen our dispositions to behave morally. Vargas can happily concede that anyone blamed under these norms would be used as a means to an end. However, such a person would also have enjoyed the benefits secured by the system's previous ascriptions of responsibility. She would have profited from the increased security that comes with living in a more moral society—a society less disposed toward acts of violence and dishonesty. Her interests would have been advanced by the reduced probability of iniquitous acts occurring and undermining her own ends. Hence, the responsibility system under which she was blamed would not have treated her as a mere tool to promote the ends of others. It would have respected her as a person whose projects ought to be protected. Because ACM promotes an end that redounds

[4] Kramer cites Jeffrey Reiman and Dan Markel as notable retributivists who have defended the deterrence-oriented rationale for capital punishment against the 'mere means' objection.

to the benefit of everyone—including those it deems blameworthy—it does not conflict with Kant's stricture in the second formulation.

Nonetheless, there are strong reasons to doubt that many deontologists would be willing to embrace ACM. One reason pertains to the restricted role that the idea of basic desert plays within Vargas's theory. This idea is captured by Derk Pereboom in the following passage:

> For an agent to be morally responsible for an action in this [the basic desert] sense is for it to be hers in such a way that she would deserve to be blamed if she understood that it was morally wrong, and she would deserve to be praised if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations. (Pereboom 2014: 2)

Given ACM's expressly forward-looking concerns, one might be surprised to learn that Vargas thinks the backward-looking notion of desert can be justified by ACM. After all, he repeatedly emphasises that we should endorse norms of responsibility that are optimal for fostering moral agency in the future. However, he also suggests this goal could be advanced if we adopted a norm of apportioning praise and blame according to basic desert (Vargas 2013: ch 8, 2015). ACM gives us 'reason to participate in retrosert [i.e. backward-looking, desert-based] practices because of the contributions of such practices and judgments in our becoming better beings' (Vargas 2015: 2668). Vargas acknowledges that this observation is unlikely to placate theorists who want their desert to be backward-looking all the way down. However, he claims that ACM at least has the virtue of providing *some* normative basis for desert: 'For those who like their desert unencumbered by larger justificatory support, [this] picture of desert… does not look like a picture of desert at all' (Vargas 2015: 2666). However, what those who attach importance to basic desert typically want is not an account that is utterly devoid of theoretical support, but an account that is not grounded in a consequentialist morality. This is encapsulated in the retributivism of Michael Moore:

> Retributivism… is the view that punishment is *justified* by the desert of the offender. The good that is achieved by punishing… has nothing to do with future states of affairs, such as the prevention of crime or the maintenance of social cohesion. Rather, the good that punishment achieves is that someone who deserves it gets it. (Moore 1997: 87, emphasis added)

Philosophers of this ilk will not be impressed by Vargas's attempt to ground basic desert in consequentialist reasoning. Of course, such theorists could well be ethically mistaken. However, this is beside the point. The key issue here is whether ACM can be readily adopted by theorists in diverse quarters. In failing to allow any justificatory role for basic desert, Vargas has already alienated a fair portion of nonconsequentialists.

Furthermore, many deontologists who do not care much for basic desert will find ACM's future-oriented concerns unpalatable. Consider, for instance, what a Kantian would have to maintain if she embraced ACM. She could happily accept that the reactive attitudes help us to internalise moral norms. However, she could not seriously claim that they make us internalise those norms for the *right reasons*. It would be incredible to think that praise and blame could improve our abilities to derive moral precepts from the Categorical Imperative. But once she has conceded this much, the Kantian's endorsement of ACM becomes vulnerable to a complaint famously made by Hegel. Offering people external incentives to comply with morality is like 'raising one's stick at a dog' to improve its behaviour (Hegel 1991: 125).

Moreover, once a deontologist has adopted this understanding of moral responsibility, she will find herself having to explain why it is appropriate to manipulate people through praise and blame but not through more rebarbative methods like brainwashing and the giving of evil-tasting pills. She might claim that these methods fail to respect the wrongdoer's right to self-directed agency in a way that praise and blame do not. For instance, when we blame someone for a wrongful act, it remains open for them to perform similar acts in the future. In contrast, such acts could be permanently foreclosed by brainwashing and mind-altering pills. This line of reasoning is, however, untenable. First of all, the reactive attitudes could produce similarly strong dispositional changes in some people. Second, it is possible to imagine milder forms of brainwashing and pill-giving that generate strong but not incapacitating aversions. These methods would not fall afoul of the self-directed agency requirement as understood here. Hence, it is not obvious that deontologists can consistently embrace ACM while repudiating the manipulative techniques just discussed.

The strongest reason to think that most deontologists would disavow ACM is that it seems to require them to renounce their deontology. Vargas invites them to adopt ACM by appealing to the utility of praise and blame in promoting compliance with deontological norms. In doing so, he is suggesting that compliance with deontological duties be treated as an intrinsic good that ought to be maximised. However, anyone who takes this maximising view is aptly characterised not as a deontologist but as what Nozick calls a 'utilitarian of rights' (Nozick 1974: 30). On this view, minimising duty violations (thereby protecting the rights which are correlative to those duties) 'merely would replace the total happiness as the relevant end state in the utilitarian structure' (Nozick 1974: 28, emphasis removed).[5] Nozick claims that this understanding would distort the function of rights in governing our actions. Rights do not specify end states to be pursued; rather, they 'express the inviolability of others' by constraining what we may do to them (Nozick 1974: 28, 29, 32). Rights 'reflect the fact that no moral balancing act can take place among us; there is

[5] In this paragraph, I take for granted the view of Wesley Newcomb Hohfeld that rights and duties are correlative. As put by Kramer, who defends this position at length, rights and duties are mutually entailing in that 'each is the other from a different perspective, in much the same way that an upward slope viewed from below is a downward slope viewed from above' (Kramer 1998: 24, 24–60). Those inclined to dissent from this view can instead see Vargas as offering a utilitarianism of *duties* that deontologists would find equally disagreeable.

no moral outweighing of one of our lives by others so as to lead to a greater overall *social* good' (Nozick 1974: 33, emphasis in original). There is no need to assess Nozick's conception of rights here. The important point is that it reflects a paradigm of deontological ethics. Any deontologist who wished to embrace ACM would have to forsake the understanding of rights that is characteristic of deontology. For some, ACM may be sufficiently alluring to induce this shift, but for most it will not. Most deontologists will wish to remain deontologists.

The majority of virtue ethicists would also find Vargas's model unsavoury, notwithstanding their concern about cultivating morally desirable characteristics. This is because modern proponents of virtue ethics usually consider virtues to involve moral understanding rather than mere temperaments.[6] An honest person, for example, is not just someone who is inclined to represent her mental states accurately. Indeed, too much of this could bring her into the realm of vice. Rather, an honest person tells the truth because she appreciates the value of doing so. As the editors of a recent collection on virtue ethics note, 'in order for a moral character trait to be a virtue, it must not only be in accord with the relevant moral norms, but the disposition must also be informed by proper reasoning about the matter at hand' (Timpe and Boyd 2014: 7). Hence, the tolerant moral drifter who appeared in the previous subsection would not be regarded as virtuous even if his behaviour happened to be exemplary. As with deontologists, then, Vargas will struggle to win over any virtue ethicists in the absence of strong evidence that praise and blame conduce to sound moral reasoning.

This section's broad line of reasoning can be captured by an autobiographical analogy. As a child, I would often wedge a piece of cardboard into my bicycle so that it intersected with the spokes on one of the wheels. When cycling, this resulted in a vibrating sound that (to the fertile imagination of a child) resembled an engine, enabling me to pretend I was riding a motorcycle. Alas, I was just riding a bicycle. This anecdote helps to underscore the role that moral agency plays in Vargas's account. He has taken a bicycle (MIT), inserted a piece of cardboard (the notion of moral agency), and is claiming to ride a motorcycle (ACM). But the sound of ACM's revving engine is sustained by nothing more than a flimsy piece of cardboard. On closer inspection, we can see that Vargas is riding the same model as his predecessors. His account does not merely share MIT's shortcomings; it just is MIT.

3 Free Will and Agency Cultivation

Having exposed Vargas's failure to advance beyond MIT, I now wish to turn my attention to his proposed test for responsible agency. Given ACM's aims, 'it only makes sense to demand adherence to norms of praise and blame if one has the capacity to regulate one's conduct in light of moral considerations' (Vargas 2013: 197). To aid our assessment of which wrongdoers possess this capacity, Vargas proposes a rather abstruse counterfactual test. This final section begins by briefly

[6] I am grateful to an anonymous reviewer for reminding me of this.

recounting his test for responsible agency, along with some of the motivations for construing it as he does. I then examine the test in greater detail and home in on its free will condition. In short, Vargas maintains that whenever some agent *S* has failed to respond to some moral consideration *M*, *S* had the free will to respond to *M* if and only if *S* would have responded to *M* in a *suitable proportion* of *relevantly similar* worlds. His guidance on how to understand the proportionality and similarity requirements leaves much to the reader's imagination, and so I begin by trying to follow his recipe for making these notions more precise. After doing so, I conclude that we can only determine whether *S* satisfies Vargas's counterfactual test if we already know whether ACM gives us good reasons to hold people like *S* responsible. But once we possess such knowledge, any enquiry into how *S* would have behaved in various alternative worlds becomes superfluous—there is no longer any need to contemplate any counterfactuals. I suggest that this defect arises because Vargas wants his account of free will to perform two quite different roles. First, he wants it to reliably pick out agents who are likely to be positively influenced by blame. Second, he wants it to capture the intuition that an agent must be able to do other than he does in order to be blameworthy.

3.1 Testing for Responsible Agency

How can we tell whether an individual possesses the capacity to act on the basis of moral considerations? Vargas's answer is located within a two-part test that sets out the criteria for responsible agency. The test consists of a self-directed agency requirement and a free will requirement, and these broadly correspond to Fischer and Ravizza's epistemic and control conditions (Fischer and Ravizza 1998). Self-directed agency involves properties like 'beliefs, desires, means-end reasoning, the ability to formulate and execute action plans, and the presence of ordinary epistemic abilities' (Vargas 2013: 213, footnote omitted). But the capacity to respond to moral considerations is located within the free will condition, and it is this condition that Vargas is primarily concerned to explicate. Before we unpack this requirement, some preliminaries are in order. First, it should be borne in mind that Vargas's conception of free will is explicitly tied to his concerns about cultivating moral agency. He wants to develop an account of free will which, if widely used to guide ascriptions of responsibility, would strongly conduce to moral considerations-responsiveness. In his own words, 'the powers that constitute free will are those that suffice to support morali[s]ed praising and blaming practices, so that such practices increase our acting on moral considerations and expand the contexts in which we do so' (Vargas 2013: 213). He emphasises this further by stating that the metaphysics of his account are 'structured throughout by the normative concerns that undergird our interest in responsibility' (Vargas 2013: 215).[7] This strategy brings to mind Dennett's approving summation of Stephen White's methodology: 'Don't try to use metaphysics to

[7] The following assertion from McGeer is therefore puzzling: 'the features that make for responsibility [on Vargas's account] are ontologically prior to, and *conceptually independent* of their role in making agents fit or appropriate targets of the responsibility system' (McGeer 2015: 2636, emphasis in original). Vargas himself has taken exception to this claim (Vargas 2015: 2676).

ground ethics… put it the other way around: Use ethics to fix what we should mean by our 'metaphysical' criterion' (White 1991: ch 8; Dennett 2003: 297).

Vargas clearly wants his test to identify people who are amenable to the reactive attitudes. However, he does not recommend that we try to calculate the consequences of each particular instance of praise and blame. Rather, he favours a set of responsibility norms that are generally efficacious. Any such system of norms will inevitably ascribe free will to people whose behaviour is immune to or even exacerbated by praise or blame (Vargas 2013: 177–181). As Vargas states:

> [T]here may be instances where my gratitude or indignation may fail to influence anyone in the proper fashion. Nonetheless, my gratitude (or indignation) can have an appropriate role, internal to a system of moral influence, because the prevalence of such attitudes and corresponding practices contributes to the efficacy and stability of the responsibility system over time. (Vargas 2013: 177, footnote omitted)

He goes on to state: 'At a general level, the norms of praise and blame just are those that are most effective at collectively influencing agents in the appropriate way' (Vargas 2013: 180). With these matters in mind, we can now turn to Vargas's two-part test for responsible agency, which takes the following form:

An agent *S* is a responsible agent with respect to considerations of type *M* in circumstances *C* if *S* possesses a suite of basic agential capacities implicated in effective self-directed agency (including, for example, beliefs, desires, intentions, instrumental reasoning, and generally reliable beliefs about the world and the consequences of action) and is also possessed of the relevant capacity for (A) detection of suitable moral considerations *M* in *C* and (B) self-governance with respect to *M* in *C*. Conditions (A) and (B) are to be understood in the following ways:

(A) The capacity for detection of the relevant moral considerations obtains when:

- (i) *S* actually detects moral considerations of type *M* in *C* that are pertinent to actions available to *S* or
- (ii) in those possible worlds where *S* is in a context relevantly similar to *C*, and moral considerations of type *M* are present in those contexts, in a suitable proportion of those worlds *S* successfully detects those considerations.

(B) The capacity for volitional control, or self-governance with respect to the relevant moral considerations *M* in circumstances *C* obtains when either

- (i) *S* is, in light of awareness of *M* in *C*, motivated to accordingly pursue courses of action for which *M* counts in favo[u]r, and to avoid courses of action disfavo[u]red by *M* or
- (ii) when *S* is not so motivated, in a suitable proportion of those worlds where *S* is in a context relevantly similar to *C*.

(a) *S* detects moral considerations of type *M*, and
(b) in virtue of detecting *M* considerations, *S* acquires the motivation to act accordingly, and
(c) *S* successfully acts accordingly (Vargas 2013: 213–214).

For Vargas, an agent has free will if and only if conditions (A) and (B) are jointly satisfied.[8] As he acknowledges, questions immediately arise about how to flesh out the proportionality and similarity requirements of (A.ii) and (B.ii). Just how similar must a possible world be to our own in order to qualify as *relevantly* similar? And what proportion of relevantly similar worlds is a *suitable* proportion? Rather than proposing that we adopt rigid understandings of these concepts, Vargas suggests that we cash them out according to 'the standards an ideal, fully informed, rational, observer in the actual world would select as at least co-optimal for the cultivation of moral considerations-responsive agency' (Vargas 2013: 214). He also tells us that 'the notion of contextual relevance will be settled by what is *actually co-optimal or better for fostering agency that recogni[s]es and suitably responds to moral considerations in the actual world, in ordinary contexts of action*' (Vargas 2013: 219, emphasis in original). Similarly, 'the suitable proportion of worlds will be that proportion—whichever it is—which is at least co-optimal for a practicable system of judgments, practices, and attitudes shaped by the aims of the responsibility system' (Vargas 2013: 220–221). In other words, we must tailor our understanding of the similarity and proportionality requirements in order to maximise the number of instances in which free will is ascribed to people who are likely to be positively influenced by blame, while minimising the number of instances in which free will is ascribed to be people for whom a positive influence is unlikely. The repeated use of the adjective 'co-optimal' is presumably intended to capture the fact that different configurations of the proportionality and similarity requirements may be combined to yield equally efficacious responsibility judgments. For instance, if we conjoined a *stringent* similarity requirement with a *loose* proportionality requirement, our ascriptions of praise and blame could be effective in developing moral agency, say, 75% of the time. Similarly, a *loose* similarity requirement conjoined with a *stringent* proportionality requirement might also produce effective results 75% of the time. If there were no possible configuration that were more efficacious, the above configurations would be co-optimal.

Of course, this talk of loose and stringent understandings is highly vague, and few would find Vargas's account persuasive unless these requirements could be

[8] It should be noted that he does not construe free will as an all-or-nothing, acontextual property. For instance, an agent may have free will with respect to M in C_1 but not in C_2. One person may generally possess the free will to respond to M_1 but not M_2, while another may generally possess the free will to respond to M_2 but not M_1. Vargas intends this *circumstantialist* picture to distance his accounts from the *atomistic* (acontextual) and *monistic* (non-scalar, all-or-nothing) theories that predominate the reasons-responsive literature. He claims that most reasons-responsive accounts are insufficiently attentive to the findings of *situationist social psychology*, which show that contextual factors have a far greater impact on human action than is ordinarily assumed. Although Vargas's circumstantialism and his criticisms of traditional reasons-responsive accounts are important contributions to the responsibility literature, I shall have very little to say about them here (see Vargas 2013: 204–209).

delineated with far greater specificity. It is therefore regrettable that he does not provide much more advice for the reader who wishes to add some meat to his skeletal framework. To be sure, he does provide *some* further advice. For instance, he cautions against an extremely fine-grained understanding of the relevance requirement whereby a context is germane only if it is identical to the real-world situation in which an agent acted. After all, if determinism were true and we adopted this understanding, then all relevant contexts would contain the same outcome as in the original situation. If *S* failed to respond to a moral consideration in the real world, he would also fail in all relevantly similar worlds. But this approach would be unpalatable. It would recommend that we never blame anyone for anything, thereby depriving us of all the benefits that blame can bring (Vargas 2013: 218–219).

Vargas also counsels against a capacious understanding of contextual relevance, which would generate too many false positives to warrant our endorsement. Such an understanding would make it exceedingly hard for even the most scrupulous people to avoid being blamed frequently (Vargas 2013: 218–221). He makes similar remarks about the proportionality requirement, telling us, for example, that no one 'in the sublunary realm' could respond to the relevant consideration(s) in all relevant worlds, whilst everyone could in at least one world (Vargas 2013: 221–222). ACM thus requires 'a Goldilocks notion of capacity: not too strict and not too lax, or, just right for enabling a well-functioning system of blame' (Vargas 2018: 120).

3.2 Epistemic Redundancy and Conceptual Distortion

Vargas's advice is quite unilluminating for those who want to concretise his conception of free will. It is therefore surprising that few of his critics have trained their attention on it. Indeed, Vargas himself acknowledges that he offers 'more of a recipe for a substantive conclusion than a bold, decisive answer' (Vargas 2013: 222). Let us follow his recipe and sample the results. Perhaps Vargas's conception of responsible agency can be given more precision by pondering how his test should apply to various classes of individuals. Consider psychopaths, for example. Before he had published his counterfactual test, Vargas wrote a piece about the responsibility of psychopathic serial killers. In it he lists various hallmarks of psychopathy: poor impulse control resulting from observable brain abnormalities, lack of empathy and guilt, no aversion to violent imagery, and so on. Interestingly, psychopaths are unable to distinguish moral norms from mere conventions. They would see no qualitative difference between a prohibition on torturing babies and a restriction on the number of players in a game of football. Vargas notes that psychopaths are notoriously unresponsive to social pressures. They are blind to the moral valence of their acts, and their inability to feel shame or contrition makes them largely unresponsive to hostile attitudes (Vargas 2010: 70–73). Thus, ACM gives us little reason to hold them responsible. They are unlikely to internalise moral norms as a result of praise and blame. In Vargas's words, 'we ought not (morally) to blame psychopaths for

harming others. That means that psychopathic serial killers aren't, properly speaking, blameworthy for what they do' (Vargas 2010: 74, endnote omitted).[9]

In light of this, Vargas would clearly want us to construe his counterfactual test in such a way that any psychopath who fails to respond to some moral consideration *M* would not have responded to *M* in a suitable proportion of relevantly similar worlds. This might allow us to say certain things about which worlds count as relevant. For instance, a world in which someone responds to *M* because he is no longer a psychopath would obviously be irrelevant. Similarly, if we assume kleptomania is not influenced by blame, ACM would give us little reason to contemplate a world in which a thief is no longer a kleptomaniac. However, these restrictions on the counterfactual analysis are fairly unenlightening and it is unclear that Vargas gives us the means to say anything more significant. Moreover, the examples just adduced help to bring out the superfluity of his test. If we endorse ACM and we know that psychopaths are incorrigible, we can simply bypass Vargas's abstruse test and adopt a norm of exempting psychopaths. The same applies *mutatis mutandis* to acts of kleptomania.[10] This point can be generalised across any number of examples.

Consider, for instance, a schoolteacher who endorses ACM and wants to apply Vargas's counterfactual test to various disruptive teenagers. After reading *Building Better Beings*, she finds herself confronted with a boisterous boy whose rowdy behaviour is intended to boost his reputation for rebellion. Noticing that he has failed to respond to the moral considerations against disturbing class, she must now ask whether he would have responded to them in a suitable portion of relevantly similar worlds. Puzzled at how to answer this question, she revisits Vargas's guidance. This reminds her that the proportionality and similarity requirements should be tailored to increase the probability that freedom is ascribed to wrongdoers who can be influenced positively by blame. She also remembers Vargas's exhortation to avoid predicting the effects of praise and blame in any given case. Her interest should not be in whether blaming *this* boy would yield positive results but whether a general policy of blaming boys *like him* would do so.[11] In order to apply the counterfactual test in this particular case, she must flesh out the proportionality and similarity requirements by reflecting on whether it is, on balance, useful to blame boisterous boys. While she cannot find much data on the matter, her experience suggests that chastisement normally works on such boys,

[9] In *Building Better Beings*, Vargas countenances 'acquired psychopathy' as a valid excuse, thereby tacitly reaffirming his view that psychopaths are exempt from blame (Vargas 2013: 278). However, this view does not sit well with his commodious notion of responsiveness to moral considerations. Even if psychopaths lack moral sentiments, many are able to adjust their conduct in light of prudential concerns. It seems plausible to think that repeated exposure to negative stimuli (e.g. blame) could give at least some psychopaths an affective drive not to engage in wrongful conduct. According to ACM, this should make them moral agents. (While poor impulse control is a *paradigmatic* trait of psychopaths, it is not a *necessary* trait. A high degree of prudence does not disqualify someone from being a psychopath—'functional psychopath' is not a contradiction in terms. If someone possesses most of the other paradigmatic features to a substantial extent, he will still qualify a psychopath.)

[10] Assuming again that kleptomania—an underexplored condition—is not diminished by blame.

[11] 'On the account that I propose, whether the agent is morally responsible for his or her actions is not a function of a particular agent's susceptibility to influence in that particular circumstance, but rather a function of what the justified norms of moral influence say about the status of responsible agents in those contexts' (Vargas 2013: 103).

perhaps because it makes them feel like fools in the eyes of their peers. She therefore resolves to construe the test accordingly: whenever faced with a boisterous boy, she will fix the requirements to yield the conclusion that he would not disrupt class in an appropriate number of germane worlds. In contrast, she believes that reprobation only exacerbates the misbehaviour of pupils with 'rotten social backgrounds' by heightening their sense of disenfranchisement. As a result, she concludes that any such pupil would not act differently in enough worlds to qualify as a responsible agent.

These reflections should leave one wondering what purpose the counterfactual test really serves. It is initially framed as a means of guiding our assignments of blame. We should only blame Jones, Vargas tells us, if we are convinced that he would have behaved differently in apposite circumstances. But we cannot determine how Jones would have behaved in the relevant circumstances until we have decided whether it is good policy to blame similar people for similar wrongdoings. If we have not settled this matter, we will not be able to give any determinate meaning to the relevant counterfactual requirements. However, once the matter is settled, we will be left with no reason to engage in any counterfactual reasoning. If it is good policy to blame people like Jones, ACM will permit us to go ahead and hold them responsible. The work necessary to give the counterfactual test meaning makes that very test redundant. Once we know that people like Jones are susceptible to blame, there is no need to agonise about how he might have behaved in various alternate universes. We do not have to sift through a multitude of hypothetical cases, assessing them for relevance and dividing them according to whether the desired behaviour occurred. To use another example, suppose that reprimanding religious bigots has desirable effects. If this is so, we can be certain that the test will ascribe free will to a run-of-the-mill religious bigot before any straining of the imagination has occurred. In contrast, if alcoholics respond negatively to blame, we know they will not satisfy the test.

Vargas's notion of free will, it turns out, resembles the whistle on an engine—while it may seem to *make* us blameworthy, it is actually our blameworthiness that makes us free. Whenever we have good forward-looking reasons to deem some class of wrongdoers blameworthy, Vargas instructs us to ascribe free will to its members and speak as though their freedom made them blameworthy all along. But contrary to appearances, questions of free will are actually settled by questions about the propriety of the reactive attitudes, rather than the other way around. A wrongdoer's blameworthiness is logically prior to his free will: we must decide whether it is good policy to blame him *before* we can ascribe counterfactual freedom to him. Indeed, the statement that he has free will is apparently just a roundabout way of saying that ACM gives us good (freedom-independent) reasons to hold him responsible. But this makes counterfactual freedom both meaningless and useless. If freedom simply means belonging to a group of people whom it is good policy to blame, it is not really freedom in any recognisable sense of the term. Moreover, since we must settle the relevant policy questions in advance of applying the counterfactual test, there is no clear sense as to how the test is supposed to aid us. In short, if we embrace ACM, enquiries into an agent's counterfactual freedom will become superfluous when considering the ethics of holding him responsible.

What has led Vargas to invert the relation between freedom and responsibility in this way? The answer lies in his desire to achieve two quite disparate goals. First, as has already been emphasised, he wants to tailor his metaphysics to his teleology. Second, he seeks to preserve the ostensible folk commitment that moral responsibility requires the ability to do otherwise. To understand this second goal, it will be helpful to reflect briefly on an aspect of Vargas's methodology that we have not yet covered. Vargas believes that the folk concept of responsibility is implicitly libertarian but that the agential powers posited by libertarians are metaphysically and scientifically implausible (Vargas 2013: ch 1, ch 2). However, he does not want to jettison the folk concept of responsibility altogether. Rather, he embraces a position of 'philosophical conservation' according to which 'we ought to abandon our standing commitments only as a last resort' (Vargas 2013: 73). Thus, Vargas prefers a *revisionist* position that seeks merely to modify moral conventions in light of deeper moral concerns (Vargas 2013: ch 3).[12]

It should be noted that there is not really anything *philosophical* about Vargas's conservation principle—it is in fact entirely *pragmatic* (in the non-philosophical sense). The principle is not based on the view that folk convictions are *prima facie* credible. Nor does it have the epistemic goal of bringing us closer to truth. Rather, it appeals entirely to the social costs involved in substantial alterations to conventional morality. Because our minds are 'cognitively finite systems with limited resources', it is enormously taxing to initiate vast changes in what we believe and how we behave. Concerning our beliefs, Vargas writes that 'the larger the revision, the more general doxastic stability is threatened and the larger the tear will be in the fabric of interlocking justification and explanation that ordinarily develops among our beliefs' (Vargas 2013: 74). Regarding behaviour, he asserts that 'in cases where the relevant beliefs are intimately connected with practical matters, where those beliefs structure our practices and interactions, the costs of belief revision are particularly high because revision disrupts entrenched dispositions of action and patterns of conduct' (Vargas 2013: 74). This may be good advice for policymakers, but calling it *philosophical* conceals its intent. Vargas's principle does not speak to philosophers concerned about *what we should do* but to practical people concerned about *the most we can expect people to do*. This is rather like the posture of an environmental campaigner who does not call for a complete overhaul of our practices because such high demands would overwhelm many people into complete inaction. While this orientation is no doubt laudable, it is a key source of Vargas's conceptual distortion.

Vargas thinks that modern Westerners possess untenable libertarian commitments, but he wants to offer them a way to preserve the idea that blame requires the ability to do otherwise (albeit in a non-libertarian sense). He claims that his account 'permits us to make sense of the idea that (at least sometimes) agents can do otherwise when acting' and 'provides a way to accommodate the familiar thought that there are responsibility-significant possibilities regarding what could have happened, but did not' (Vargas 2013: 230). Given Vargas's concern about our cognitive

[12] While Vargas offers a rich discussion of revisionism and the various forms it may take, there is no need to recapitulate it here.

limitations, it is ironic that he prescribes an abstruse counterfactual test to guide everyday ascriptions of responsibility. Equally troubling, however, is that his construal of the counterfactual test calls for a much more profound shift in folk thinking than he recognises. On Vargas's understanding, whenever *S* has failed to respond to *M* in *C*, the statement '*S* would have responded to *M* in *C* in a suitable proportion of relevantly similar worlds' is really a convoluted way of saying something like 'blaming people such as *S* helps indispose them toward future misconduct'. As previously noted, if ACM gives us good reasons to blame a class of people, we are told to ascribe certain metaphysical properties to members of that class and then claim that our reasons emanated from those metaphysical properties all along. Thus, contrary to what nearly everyone thinks, free will is no longer a prerequisite of justified blame; justified blame is in fact a prerequisite of free will. Anyone who fully appreciates Vargas's tortuous counterfactual test will regard it as a quagmire of evasion at best or a wretched subterfuge at worst.

4 Concluding Remarks

The ostensible superiority of Vargas's account is sustained by the putative contrast between ACM's focus on moral agency and MIT's focus on being disposed to act morally. However, I have argued that his conception of agency is so expansive that it cannot be distinguished from a disposition to behave in a morally desirable manner. ACM retains all of MIT's shortcomings because it just is MIT under a different name. Vargas could always adopt a more refined account of moral agency, but doing so would increase the evidential burden on his empirical claims about how the reactive attitudes influence us. While his counterfactual test is intended to preserve folk views about the relationship between freedom and responsibility, it ends up distorting them by tacitly making questions of just blame logically prior to questions of free will. This distortion results from an attempt to tailor his metaphysics to his ethics while preserving the everyday intuition that alternative possibilities are necessary for freedom. Regrettably, the pursuit of two divergent goals has resulted in the achievement of neither.

Acknowledgements Many thanks to Matt Kramer for valuable comments and also to Antje du Bois-Pedain and Findlay Stark for their searching queries.

References

Arneson, R. 2003. The smart theory of moral responsibility and desert. In *Desert and justice*, ed. Serena Olsaretti, 233–263. Oxford: Oxford University Press.

Bartal, I.B., J. Decety, and P. Mason. 2011. Empathy and pro-social behaviour in rats. *Science* 344: 1427–1430.

Boonin, D. 2008. *The problem of punishment*. New York: Cambridge University Press.

Campbell, C.A. 1951. Is 'freewill' a pseudo-problem? *Mind* 60: 441–465.

Capes, J.A. 2016. Review of building better beings: A theory of responsibility, by manuel vargas. *Journal of Moral Philosophy* 13: 245–248.

Churchland, P. 2019. *Conscience*. New York: Norton.

Dennett, D.C. 2003. *Freedom evolves*. London: Penguin Books.

Dennett, D.C. 2015. *Elbow room: The varieties of free will worth wanting*, New ed. Cambridge: MIT Press.

Dolinko, D. 1999. Morris on paternalism and punishment. *Law and Philosophy* 18: 345–361.

Dworkin, G. 1986. Review of Elbow Room, by Daniel C Dennett. *Ethics* 96: 423–425.

Dworkin, R. 2011. *Justice for hedgehogs*. Cambridge: Harvard University Press.

Elzein, N. 2013. Basic desert, conceptual revision, and moral justification. *Philosophical Explorations* 16: 212–225.

Fischer, J.M., and M. Ravizza. 1998. *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Goodall, J. 2010. *Through a window: My thirty years with the chimpanzees of Gombe*. New York: Mariner Books.

Hegel, G.F.W. 1991 *Elements of the philosophy of right*. Translated by HB Nisbet. Cambridge: Cambridge University Press.

Kant, I. 1998. *Groundwork of the metaphysics of morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press.

Kramer, M.H. 1998. Rights without trimmings. In *A debate over rights*, ed. Matthew Kramer, N.E. Simmonds, and Hillel Steiner, 7–111. Oxford: Oxford University Press.

Kramer, M.H. 2011. *The ethics of capital punishment*. Oxford: Oxford University Press.

McGeer, V. 2015. Building a better theory of responsibility. *Philosophical Studies* 172: 2635–2649.

McKenna, M. 2012. *Conversation and responsibility*. Oxford: Oxford University Press.

Moore, M.S. 1997. *Placing blame: A theory of criminal law*. Oxford: Oxford University Press.

Morris, H. 1981. A paternalistic theory of punishment. *American Philosophical Quarterly* 18: 263–271.

Nelkin, D.K. 2011. *Making sense of freedom and responsibility*. Oxford: Oxford University Press.

Nowell-Smith, P. 1948. Freewill and moral responsibility. *Mind* 57: 45–61.

Nozick, R. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Pereboom, D. 2014. *Free will, agency, and meaning in life*. Oxford: Oxford University Press.

Schlick, M. 1962. *Problems of ethics*. New York: Dover Publications.

Smart, J.J.C. 1961. Free-will, praise and blame. *Mind* 48: 291–306.

Sripada, C., and S. Stich. 2006. A framework for the psychology of norms. In *The innate mind: Volume 2: culture and cognition*, ed. Peter Carruthers, Stephen Laurence, and Stephen Stich, 280–301. Oxford: Oxford University Press.

Strawson, P.F. 1993. Freedom and resentment. In *Perspectives on moral responsibility*, ed. John M. Fischer and Mark Ravizza, 45–66. New York: Cornell University Press.

Timpe, K., and C.A. Boyd. 2014. *Virtues and their vices*. Oxford: Oxford University Press.

Vargas, M. 2010. Are psychopathic serial killers evil? Are they blameworthy for what they do? In *Serial killers—Philosophy for everyone: Being and killing*, ed. S. Waller, 66–77. Oxford: Wiley-Blackwell.

Vargas, M. 2013. *Building better beings: A theory of moral responsibility*. Oxford: Oxford University Press.

Vargas, M. 2015. Desert, responsibility, and justification: a reply to Doris, McGeer, and Robinson. *Philosophical Studies* 172: 2659–2678.

Vargas, M. 2018. The social constitution of agency and responsibility. In *Social dimensions of moral responsibility*, ed. Katrina Hutchison, Catriona Mackenzie, and Marina Oshana, 110–136. Oxford: Oxford University Press.

Wallace, R.J. 1994. *Responsibility and the moral sentiments*. Cambridge: Harvard University Press.

White, S. 1991. *The unity of self*. Cambridge: MIT Press.

Wolf, S. 1990. *Freedom within reason*. Oxford: Oxford University Press.
Zahn-Waxler, C., M. Radke-Yarrow, and R.A. King. 1979. Child rearing and children's prosocial initiations toward victims of distress. *Child Development* 50: 319–330.